Mit iszunk? A Magyar WordNet automatikus kiterjesztése szelekciós preferenciákat ábrázoló szófajközi relációkkal

Miháltz Márton, Sass Bálint

MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport, 1444 Budapest, Pf. 278.
{mihaltz.marton, sass.balint}@itk.ppke.hu

Kivonat: A cikkben bemutatott, folyamatban lévő munkálatok célja a Magyar WordNet automatikus kiegészítése új, különböző argumentumpozíciók szelekciós preferenciáit ábrázoló ige-főnév relációkkal. Bemutatunk egy algoritmust, amely korpuszgyakorisági adatok és a WordNet hierarchikus szerkezete alapján megkísérli azonosítani a vonzatpozíciók szemantikai típusait legjobban reprezentáló HuWN hipernima-algráfokat. Az eljárás segítségével minden, a korpuszban megtalálható, esetraggal vagy névutóval jelölt igei argumentumpozíciót igyekszünk lefedni. Nem célunk egyértelmű, kizárólagos kategóriák kijelölése, ehelyett súlyozott listák segítségével igyekszünk felsorolni a megfigyelt példákból általánosítható leggyakoribb típusokat. Az eredmények reményeink szerint a Magyar WordNet felhasználóin felül az általunk fejlesztett szintaktikai elemző számára is hasznos erőforrásként fognak szolgálni. A cikkben bemutatunk néhány előzetes eredményt és szót ejtünk néhány felmerülő kérdésről.

1 Bevezetés

1985-ös első kiadása óta a Princeton Wordnet (PWN) [5] mára általánosan elterjedt lexikális szemantikai erőforrássá vált a nyelvtechnológiai kutatásokban és alkalmazásokban. Szabad hozzáférhetősége, tekintélyes lefedettsége és folyamatos fejlődése mind hozzájárultak sikereihez.

Története során több lehetséges irány megfogalmazódott a PWN további javíthatósága szempontjából. Az NLP-felhasználók szemszögéből a PWN egyik hiányossága, hogy a szófajokon belül meglévő gazdag relációrendszerhez képest jóval kevesebb szófajok közötti (különböző szófajú synseteket összekapcsoló) relációt tartalmaz. A főnevek, igék, melléknevek és határozószók alhálózatai között jelenleg csak morfológiai (derivációs) kapcsolatok vannak, pl. *research* (ige) — *researcher* (fn), *engage* (ige) — *engagement* (fn) stb.

Jelen kutatás célja, hogy automatikus módszereket találjunk arra, hogy a Magyar WordNetet (HuWN) [9] bizonyos, az igéket és főneveket összekötő relációkkal egészítsük ki korpuszadatok alapján. E relációk az igék és az ige mellett megjelenő adott esetragú/névutójú bővítmények között hoznak létre kapcsolatot úgy, hogy megadják a szóban forgó vonzat szemantikai típusának általánosítását legjobban reprezentáló főnévi WordNet csomópontot, pl. *{eszik}-{étel}, {ír}-{írásmű}* stb. Ez az információ

többek között felhasználható jelenleg folyó, pszicholingvisztikai relevanciájú nyelvi elemző fejlesztését célzó projektünkben is (ld. [10] és Prószéky et al jelen kötetben).

A cikk további felépítése a következő: a következő részben röviden érintjük a magyar igei argumentumszerkezet szintaxisának és szemantikájának néhány releváns kérdését, majd ismertetjük kutatásaink céljait. A 3. részben bemutatjuk a vonatkozó irodalmat, a 4. részben az általunk javasolt algoritmust, majd az 5. részben néhány előzetes eredményt. Végül ismertetjük a további lehetséges fejleszési irányokat.

2 Háttér

A magyarban az igei argumentumokat (komplemenseket) szintaktikailag az esetragok, illetve a névutók adják meg. Ezek a relációk függvényei az egyes igék vonzatkereteinek: különböző igei vonzatkeretek különböző morfoszintaktikai pozícióihoz különböző névszói fogalmak tartozhatnak (pl. *figyel valami*RE*, elkezdődik valami*Ø*, odaéget valami*T*, érdeklődik valami* UTÁN stb.)

Másfelől ez a kötődés széles spektrumot mutathat: az egyik véglet az olyan idiomatikus, nem-kompozícionális ige–igei módosító kapcsolatoké, mint pl. *hangot ad (valaminek), issza a szavát, napvilágra hoz, tenyerén hordoz* stb. A másik végletet az olyan vonzatok képviselik, amelyeknek megfeleltethetők – egy vagy több – olyan szemantikai osztállyal, amelyek produktívan képesek az adott pozícióban elfogadható kifejezések szemantikai kategóriáját megjósolni (szelekciós preferenciák): *eszik valamit {étel, ennivaló}, ír valamit {írás, írásmű}, kiönt valami {víz, víztömeg}* stb.

Gyakran egy adott ige adott vonzatpozíciójához több szemantikai kategória is tartozik, pl. *iszik valamit {folyadék: víz, sör, bor, tej, ...}* | *{becsült mennyiség: pohár, csepp, korty, ...}*. Ezek a kapcsolatok a vonatkozó kategóriákba tartozó elemek gyakoriságainak függvényében eltérő mértékű asszociációt fejezhetnek ki az ige és a fogalomosztály között.

Az alábbiakban bemutatott módszerekkel megkíséreljük a különböző argumentumpozíciókra jellemző szemantikai kategóriákat korpuszadatok alapján automatikusan megtalálni, és ezeket a Magyar WordNetben új ige-főnév relációkkal ábrázolni. Az új relációtípus minden példányához két tulajdonságot szeretnénk társítani: egyrészt a vonzatpozíciót leíró morfoszintaktikai megkötéseket (esetrag vagy névutó), másrészt a a korpuszban mért adatok kiszámított kapcsolati erősségét, melynek célja az azonos pozícióban megadható szemantikai osztályok egymáshoz képesti szerepének számszerűsítése. Például az *{iszik}-[case=ACC, p=0,8]-{folyadék}, {iszik}-[case=ACC, p=0,2]-{becsült mennyiség}* két olyan kapcsolatot jelöl, amely az *iszik* ige két, tárgyesetű vonzatpozíciójában megfigyelt szemantikai kategóriát ad meg. A *{folyadék}* és a *{becsült mennyiség}* synsetek itt önmagukon kívül összes indirekt hiponimáikat is reprezentálják, így megadnak egy-egy fogalomosztályt.

3 Kapcsolódó munka

A szelekciós preferenciák feltérképezése kulcsfontosságú az írott nyelv szemantikai feldolgozása szempontjából. A vizsgálatok célja annak megállapítása, hogy milyen szójelentések gyakoriak és/vagy megengedettek bizonyos szavak adott szintaktikai környezetében. Resnik [12, 13] munkáját követve több tanulmány is a WordNetre támaszkodott a szelekciós preferenciák megállapításában ([2, 3, 22]).

Míg az utóbbi időkben ismertetett megközelítések a Latent Dirichlet Allocation (LDA) módszerekre koncentáltak ([15, 6, 14]), az általunk bemutatott kísérlet [13]-hoz áll közelebb. A magyar nyelv esetében elsőként kíséreljük meg az igék szelekciós tulajdonságainak automatikus feltérképezését. Munkánk nem csupán az ige-tárgy (direct object) viszony szelekciós megkötésének klasszikus problémájával foglalkozik, hanem figyelembe vesszük az összes lehetséges szintaktikai argumentumtípust is (20 fölött szám a magyarban), [1] javaslatával összhangban.

Szemben azokkal a megközelítésekkel, melyek célja csupán adott argumentumszerepben előforduló szavak halmazának azonosítása (pl. [4, 17, 14]), a [13] által felvázolt és [6] által is követett iránynak megfelelően kutatásunk célja szemantikus osztályok (típusok) címkéinek hozzárendelése az argumentumpozíciókhoz. Ezt a rendelkezésünkre álló legnagyobb kiterjedésű magyar nyelvű nyelvi ontológia, a Magyar Wordnet fogalmai csomópontjainak és taxonómiai relációinak felhasználásával szándékozunk megvalósítani.

4 Módszerek

A feladat megoldására általunk alkalmazott eljárás bemenete egy szóhalmaz (egy adott ige mellett adott bővítménypozíciójában előforduló főnevek gyakorisági listája), kimenete pedig e bővítményeket reprezentáló (általánosító) HuWN synsetek súlyozott, rendezett listája. Mindegyik kimenő synset a belőle kiinduló, hiponima-relációval alkotott algráfot reprezentálja. A kimenő synseteknek az alábbi feltételeket kell minél teljesebb mértékben kielégíteniük:

Lefedettség: a synset, illetve hiponima-leszármazottai tartalmazzanak minél többet a korpuszbeli szavak közül.

Sűrűség: a synsetből kiinduló algráf minél kevesebbet olyan szót tartalmazzon, ami nincs benne az input szólistában.

Használható általánosítások: a synset és a belőle kiinduló hiponima-algráf fejezze ki az argumentpozícióba tartozó korpuszszavak jelentéseinek általánosítását, de ne legyen túl általános. Például, kevés haszna van, ha minden igei argumentumhoz az *{entitás}* fogalmat társítjuk, mivel keveset mond az egyes argumentumok szemantikai preferenciáinak sajátosságairól.

Automatikus jelentés-egyértelműsítés: ha egy igei argumentumként szereplő szónak a WN-ben több jelentése van (több synsetbe is tartozik), elvárjuk, hogy az algoritmus csak a releváns jelentés(ek) általánosításához tartozó kapcsolato(ka)t generálja. Például, az *iszik* tárgyaként előforduló *kávé* főnév két jelentésének hiponimái

közül ne a *{termés, gyümölcs}*, hanem az *{ital, italféle}* felé konvergáljon az általánosítás.

A fenti feltételek alapján javasolt algoritmusunk vázlatosan az alábbi lépésekből áll:

1. Először megkeressük az összes lehetséges synsetet, amik az input szavakat tartalmazzák (azok összes lehetséges jelentéseit), majd ezekből generáljuk a lehető leghosszabb, hipernima-reláció szerinti útvonalakat a WN gyökércsomópontjaiig. Minden, ezeken az útvonalakban bárhol szereplő csomópont (synset) a továbbiakban szemantikaiosztály-**jelölt** lesz.
2. Ezt követi a jelöltek **szűrése**: elvetjük azokat a jelölteket, amelyek csak egyetlen egy korpuszszót reprezentálnak és a korpuszszót tartalmazó synset (direkt vagy indirekt) hipernimái. Ezzel a lépéssel kiszűrjük az általánosítást nem hordozó jelölteket.
3. A következő lépésben pontozzuk a fennmaradó jelölteteket két tényező figyelembevételével: hány bemeneti szót **fednek le** és milyen **sűrű** a jelölt által megadott részgráf a bemeneti szavakra nézve (a részgráf által lefedett bemeneti szavakat tartalmazó synsetek számának és a részgráf csomópontjai számának hányadosa). Az alábbi képlettel határozzuk meg *c* synset-jelölt pontszámát (ahol *subgr(c)* a *c*-ből kiinduló hiponima-részgráf, I_c a *subgr(c)* által lefedett bemeneti korpuszszavak halmaza):

$$Score(c) = |I_c| \cdot \frac{|\{s \in subgr(c) : w \in s, w \in I_c\}|}{|subgr(c)|}$$

4. A pontozás alapján rangsorolt jelöltek közül az **N legjobbat** adjuk vissza. Ezen a ponton történhet a bemeneti szavak jelentés-egyértelműsítése: ha az N legjobb synset között van legalább kettő, ami ugyanannak a bemeneti szónak eltérő jelentéseit fedi le, akkor a (leg)magasabb ponttal bíró jelöltet tartjuk meg, a többit elvetjük. Ezt addig ismételjük, amíg nem marad több többértelműség.

A HuWN-be ezután felvehetjük az új relációkat, amelyekben az igei synseteket összekötjük a nyertes főnévi synsetekkel. A kérdéses vonzatra vonatkozó morfoszintaktikai információn felül megadjuk a kapcsolat erősségét is, melyet a lefedett szavak korpuszgyakoriságai alapján adhatunk meg (ld. 6. rész).

Az algoritmus futtatásához felhasználtuk a *Mazsola* igei bővítménytár [16] adatbázisait. A Mazsola a 187 millió szavas Magyar Nemzeti Szövegtár [20] alapján készült, 20,24 millió tagmondatban azonosították a finit igéket és az igei bővítményekként funkcionáló főnévi csoportokat, majd ezeket csoportosították szintaktikai jellemzők (esetrag, névutó) szerint.

Annak eldöntésére, hogy milyen igéknek milyen vonzatai vannak, felhasználtuk a *MetaMorpho* magyar-angol fordítóprogram szintaktikai elemzőjében [11] használt igei vonzatkeret-leíró adatbázis anyagát is. Az adatbázis több mint 18 ezer igéhez 33 ezer vonzatkeret-leírást tartalmaz, melyek megadják az adott jelentésben szereplő lehetséges vonzatpozíciókat és az azokra érvényes, attribútumokkal kifejezett lexikai, morfológiai és szintaktikai megkötéseket. A Magyar WordNet fejlesztése során az igei synsetekhez hozzárendelték ebből az adatbázisból a megfelelő vonzatkeret-

leírásokat is [9]. Ez az információ felhasználható az új ige-főnév relációk létrehozásakor az igei synsetek egyértelmű kijelölésében.

A fentiek segítségével 25 500 különböző igei vonzatkeret 32 000 lehetséges argumentumpozíciójához készítettünk szógyakorisági listákat, melyeken futtatni tudtuk szelekciós preferenciákat általánosító algoritmusunkat.

5 Eredmények

Mivel jelenleg még dolgozunk egy olyan kiértékelési módszertanon, melynek segítségével az algoritmus eredményét humán annotátorok ítéleteivel tudnánk összevetni, eredményeink szemléltetésére bemutatunk néhány kiragadott példát.

Az 1. táblázatban felsoroltunk 6 kiválasztott igei vonzatpozíciót és az algoritmusunk segítségével hozzájuk rendelt, legnagyobb ponttal rendelkező szemantikai osztályt (HuWN synseteket).

1. táblázat: Automatikusan azonosított szemantikai osztályok az igevonzatokhoz

Ige és vonzatpozíció	Szemantikai kategória
iszik ACC	{folyadék}
kigombol ACC	{ruha}
olvas ACC	{könyv}
ül SUP	{ülőbútor}
vádol INS	{bűncselekmény}
megold ACC	{nehézség}

A 2. táblázatban bemutatjuk az *iszik* ige tárgyesetű vonzatpozíciójához tartozó 5 legmagasabb pontot elérő szemantikai kategóriát, valamint ezek pontszámát, a lefedett korpuszszavak számát (c) és a kategória kiszámított sűrűségét (d).

2. táblázat: Az *iszik* tárgyesetű vonzatpozíciójához rendelt 5 legmagasabb pontot elérő synset

Pont	Szemantikai kategória	Lefedettség	Sűrűség
9,1	{folyadék}	26	0,35
8,796	{ital, italféle, italféleség}	25	0,351
4,888	{szeszes ital, szesz, ital, alkohol}	16	0,305
4,375	{rövidital, tömény ital, tömény szesz, tömény}	7	0,625
3,759	{táplálék, tápanyag}	28	0,134

A HuWN hierarchiáját megvizsgálva észrevehetjük, hogy a *{folyadék}* csomópont hipernimája az *{ital, italféle, italféleség}* fogalomnak, amely viszont hipernimája a *{szeszes ital, szesz, ital, alkohol}* synsetnek. Felmerül a kérdés, hogy ezek közül melyikhez (melyekhez) szeretnénk az *{iszik}* igei fogalmat (accusativusi minősítésű kapcsolattal) hozzárendelni? Ha a legáltalánosabb és legtöbb pontot szerzett fogalmat preferáljuk, akkor a *{folyadék}* synsetre esik a választásunk. Egy másik nézőpontból viszont az *{ital, italféle, italféleség}* relevánsabb lehet, hiszen nem minden folyadék alkalmas emberi fogyasztásra. Bizonyos alkalmazásokban viszont fontos információ

lehet a korpuszadatok tanúsága szerint a *{szeszes ital, szesz, ital, alkohol}* fogalommal megjelenő erős kapcsolat is. Azáltal, hogy meghagyjuk az *N* legmagasabb pontot elérő szemantikai kategóriát minden argumentumpozícióban, valamint ábrázoljuk ezek relatív asszociációs erősségét is, szándékaink szerint a létrehozott erőforrás jövőbeli felhasználói számára biztosítjuk a lehetőséget arra, hogy céljaik és igényeik szerint maguk hozhassák meg ezeket a döntéseket.

6 További munka

Jelenleg módszereink továbbfejlesztésén dolgozunk. Amint elérhetővé válik egy kiértékelési metodológia, lehetséges lesz a jelölteket pontozó formula további finomhangolása, valamint kísérletezhetünk a kapcsolati erősségek beállításának optimális módjával is. További, felhasználható információk a bemeneti szavak korpuszbeli gyakoriságai, a jelölt synsetek mélysége a HuWN hálózatában és az átlagos távolságok a jelölt algráfokban.

Amint láttuk, a fent vázolt megközelítésben ige-vonzat párokhoz rendeltük hozzá az abban a pozícióban előforduló főnevek listáját, és az alapján határoztuk meg a szemantikai preferenciákat leíró legvalószínűbb HuWN synseteket. Az igék bővítményei azonban kölcsönhatásban vannak egymással: gyakran előfordul, hogy az egyik bővítmény megkötése (adott szóval való kitöltése) esetén egy másik bővítményben egy speciális (csak az első bővítményben lévő szóra jellemző) szelekciós preferenciával találkozunk. Ilyen például az 'ad -t' esetén a 'hírt ad' -rÓl bővítménye, vagy a 'húz -t' esetén a 'hasznot húz' -bÓl bővítménye. Ahogy azt [19] is hangsúlyozza, fontosnak tartjuk, hogy továbblépjünk a több bővítményt egyszerre kezelni tudó modellek felé, melyek képesek felismerni a 'hírt ad', 'hasznot húz' stb. összetett egységeket és ezek argumentumainak szelekciós preferenciáit.

Mechura [8] szerint a WordNetben található egységek nem teljesen felelnek meg a szelekciós preferenciák által megkívánt egységeknek, és felteszi a kérdést: hogyan kellene egy ontológiának kinézni ahhoz, hogy a szelekciós preferenciákban szerepet játszó szemantikai típusokat pontosan tudja ábrázolni? Az algoritmusunk segítségével előállított kategóriák vizsgálata elvezethet a válaszhoz.

7 Összefoglalás

A tanulmányban bemutattunk egy módszert a Magyar WordNet automatikus kiegészítésére új, szelekciós preferenciákat ábrázoló relációkkal, ami hasznos lehet a szövegfeldolgozó alkalmazások számára. Eredményeink érdekesek lehetnek a pszicholingvisztikai kutatások szempontjából is, mivel betekintést nyújthatnak a mentális lexikon szófajközi viszonyaiba.

Köszönetnyilvánítás

Köszönjük a TÁMOP-4.2.1.B – 11/2/KMR-2011–0002 és a TÁMOP: 4.2.2/B – 10/1–2010–0014 projektek részleges támogatását.

Hivatkozások

1. Brockmann, C., Lapata, M.: Evaluating and combining approaches to selectional preference acquisition. In: Proceedings of EACL (2003) 27–34
2. Calvo, H., Gelbukh, A., Kilgarriff, A.: Distributional Thesaurus vs. WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. In: Proceedings of CI-CLing (2005) 177–188
3. Clark, S., Weir, D.: Class-Based Probability Estimation Using a Semantic Hierarchy. In: Computational Linguistics 28:2 (2002) 187–206
4. Erk, K.: A simple, similarity-based model for selectional preferences. In: Proceedings of ACL (2007) 216–223
5. Fellbaum, C. (szerk.): WordNet: An Electronic Lexical Database. MIT Press: Cambridge (1998)
6. Guo, W., Diab, M.: Improving Lexical Semantics for Sentential Semantics: Modeling Selectional Preference and Similar Words in a Latent Variable Model. In: Proceedings of NAACL-HLT (2013) 739–745
7. Kuti, J., Varasdi, K., Gyarmati, Á., Vajda, P.: Language Independent and Language Dependent Innovations in the Hungarian WordNet. In: Proc. of The Fourth Global WordNet Conference, Szeged, Hungary (2008) 254–268
8. Mechura, M.B.: What WordNet does not know about selectional preferences. In: Dykstra, A., Schoonheim, T. (szerk.): Proceedings of the 14th Euralex International Congress, Ljouwert/Leeuwarden: Fryske Akademy (2010) 431–436
9. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (szerk.) Proceedings of The Fourth Global WordNet Conference. Szeged: University of Szeged (2008) 311–321
10. Prószéky, G.: Kutatások egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozás irányában. In: Ladányi, M., Vladár, Zs. (szerk.) A XI. MANYE-konferencia előadásai (megjelenés alatt)
11. Prószéky, G., Tihanyi, L., Ugray, G.: Moose: a robust high-performance parser and generator. In: Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta: Foundation for International Studies (2004) 138–142
12. Resnik, P.: Selectional constraints: an information-theoretic model and its computational realization. Cognition 61 (1996) 127–159
13. Resnik, P.: WordNet and Class-Based Probabilities. In: Fellbaum (1998a)
14. Rink, B., Harabagiu, S.: The Impact of Selectional Preference Agreement on Semantic Relational Similarity. In: Proceedings of International Conference on Computational Semantics (IWCS) (2013)
15. Ritter, A., Mausam, Etzioni, O.: A latent dirichlet allocation method for selectional preferences. In: Proceedings of ACL (2010) 424–434
16. Sass, B.: The Verb Argument Browser. In: Sojka, P., Horák, A., Kopecek, I., Pala, K. (szerk.): 11th International Conference on Text, Speech and Dialog (TSD), Brno, Czech Republic. Lecture Notes in Computer Science 5246 (2008) 187–192

17. Tian, Z., Xiang, H., Liu, Z., Zheng, Q.: A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation. In: Proceedings of ACL (2013) 1169–1179
18. Tufiş, D., Cristea, D., Stamou, S.: Balka-Net: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, 7(1⁻2) (2004) 3–4
19. van de Cruys, T.: A non-negative tensor factorization model for selectional preference induction. In: Natural Language Engineering 16(4) (2010) 417–437
20. Váradi, T.: The Hungarian National Corpus. In: Zampolli, A. (szerk.) Proceedings of the Second International Conference on Language Resources and Evaluation. Las Palmas: ELRA (2002) 385–389
21. Vossen, P.: EuroWordNet General Document, Version 3. University of Amsterdam (1999)
22. Ye, P.: Selectional Preference Based Verb Sense Disambiguation Using WordNet. In: Proceedings of the Australasian Language Technology Workshop (2004)